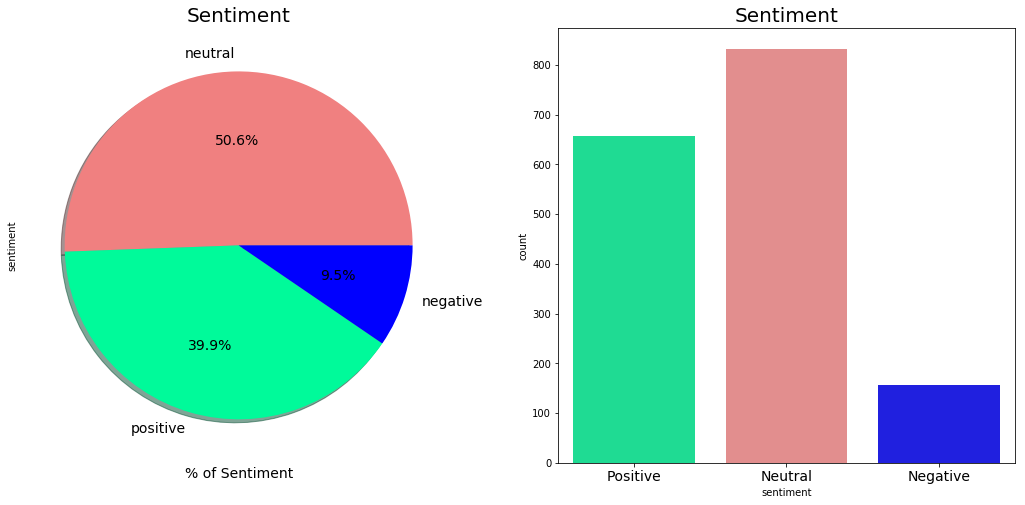
**Experiment 1**

After compiling all the json format for latest earnings call transcript paragraph/sentiment of 12 stocks we found the sentiment to be distributed in following format.



The data imbalance was clear from the above plots and thus decided to tackle this issue while splitting data into 80:20 ratio.

Before moving ahead with creating Bag of Words model in keras, we decided to try some MLP techniques to determine the accuracy and understand the flow. After using the CountVectorizer and Multinomial Naïve Bayes algorithm over sklearn’s pipeline we were able to get an accuracy of 0.537. And the confusion matrix created of the form:

array([[ 50, 57, 18],

[ 3, 604, 58],

[ 2, 167, 355]], dtype=int64)

The model was okay and further convinced us to make it better and thus we decided to try BoW + TFIDF models together and that increased the score slightly to 0.553 with confusion matrix:

array([[ 1, 96, 28],

[ 0, 632, 33],

[ 0, 214, 310]], dtype=int64)

And further experiments with BoW plus Logistic Regression gave the model a score of 0.4813 which wasn’t better than the previous two models.

Moving ahead on using Tokenizer in BoW keras model and including encoder to convert label strings to numbered index. The sequential model is tested and the model gave a best accuracy of about 0.468 after 5-7 experiments with confusion matrix of

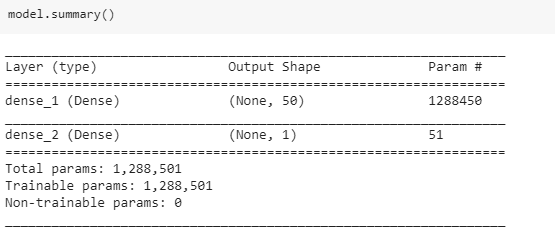
[[ 5 21 6]

[32 82 52]

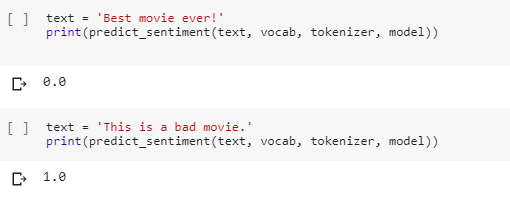
[21 84 26]]

**Experiment 2**

After training model over imdb dataset and changing maxlen right from 20 to 100, a best accuracy of 91% is achieved.



This model gave fair result when tested on imdb dataset. By giving 0 to positive review and a 1 to negative.



Eventually after testing the best model over Edgar dataset a poor accuracy of 0.19 is achieved making the prediction negative and biased for most of the cases. The confusion matrix created for this result wrt to original dataset is

[[ 0 655]

[ 0 157]]